DOCUMENT RESUME

ED 295 967 TM 011 753

AUTHOR McLarty, Joyce R.; And Others
TITLE Effects of Item Wording on Sex Bias.
PUB DATE Apr 88
NOTE 58p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988). Some pages contain small, light type.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Difficulty Level; English Instruction; *Item Analysis; Mathematics Tests; *Sex Bias; Sex Differences; *Test Construction; Test Items
IDENTIFIERS *Item Wording

ABSTRACT

The effects of superficial gender-related item wording changes on the performance of male and female examinees were studied through mathematics; discrete English items; and an English passage created in neuter, male, and female gender versions. Units of items were administered to randomly equivalent samples of about 250 examinees taking American College Testing examinations. A log-linear analysis was conducted using four factors: (1) item gender; (2) examinee sex; (3) examinee ability; and (4) response. No items were found to have sex bias, which is interpreted as the interaction of item gender with examinee sex. Mathematics items did not have varying difficulty or discrimination in the different gender versions. English items, while without varying discrimination levels in the different versions, had a dissimilar difficulty level in the gender-related versions. These differences were not systematic. Eight tables are presented, and four appendices provide the item versions. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Joyce R. McLarty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED295967

Effects of Item Wording on Sex Bias

Joyce R. McLarty

Candace Noble

Renee Huntley

American College Testing

Paper presented to the annual meeting of the National Council on Measurement in Education New Orleans, 1988.

TM011753

Effects of Item Wording on Sex Bias

Abstract

This study examined the effects of superficial gender-related item-wording changes on the performance of male and female examinees. Sixteen mathematics word problems, 21 discrete English language items, and an English passage with 28 dependent items were created in three parallel gender versions: neuter, male, and female. The discrete English items and the mathematics items were arranged into English and mathematics units, respectively, so that each item appeared in two contexts: once in a single-gender unit (e.g., all male items) and once in a mixed-gender unit. Only one version of an item appeared in a given unit. Passage-dependent items formed only single-gender units.

Units were administered to randomly equivalent samples of about 250 examinees. Loglinear analysis was used to assess the impact of item gender and its interaction with examinee sex on the difficulty and discrimination of each item in each context. The interaction of item gender with examinee sex was interpreted as an indicator of sex bias. No items were found to have sex bias in either context. Mathematics items did not have different difficulty discrimination or in the different gender versions. Neither mathematics nor English items had different discrimination levels in the different gender-related versions. English items, however, were found to have different difficulty levels in the different gender-related versions. Difficulty differences were not systematic: none of the three gender versions (neuter, male, or female) appeared consistently more or less difficult than the others.

BEST COPY AVAILABLE

Acknowledgement

The authors gratefully acknowledge the assistance of Dr. James Carlson in designing and conducting the data analysis for this study.

Effects of Item Wording

on Sex Bias

Where does test sex bias come from? Do superficial changes in the wording of test items actually make a difference in examinees' scores? Specifically, what would cause a test item to reflect a score difference between male and female examinees that is not related to the examinees' ability in the subject being tested? Take these two items, for example:

A. A carpenter has a square piece of wood that measures 60 inches on each side. He wants to make a circular table top. What is the diameter of the largest circle he can cut from this piece of wood?

B. A seamstress has a square piece of cloth that measures 60 inches on each side. She wants to make a circular tablecloth. What is the diameter of the largest circle she can cut from this piece of cloth?

Does it really make a difference whether an item that is clearly intended to measure something else (how large a circle can be cut from a square of a specified size) refers to a male or female person ("he", "she") or traditionally male or female occupation ("carpenter", "seamstress")? Some researchers have suggested that this is the case.

Numerous studies have been conducted in recent years to determine sources of sex bias in testing. Significant documentation exists of differential performance by males and females, particularly on quantitative measures of performance. Of concern are three basic issues: (1) to reduce, insofar as possible, any unfairness in test items that may exist, (2) to identify true sources of sex performance differences if they exist, and (3) to identify sources of sex performance differences that spring from the measurement of

extraneous content (Boldt, 1983). It is to the third issue that this paper is addressed, with the additional expectation that the results of this study may shed light on the equally critical issues of fairness in testing and the identification of true sources of sex performance differences.

References to males and females directly, by name, or indirectly, by pronoun, as well as references to so-called male and female values, roles, and occupations, have been considered as potential sources of differential item performance. The measurement community in general and the test publishing industry in particular are addressing the notion that such references may have some effect on test performance. Steps are being taken to reduce bias in test items in those cases where certain aspects of the content or context of an item are found to have a negative effect on the performance of a given group. Test constructors generally agree that items that are demeaning to any group or that portray a group in a stereotypic manner should not be included in a test, even if it is shown that such items have no impact on test performance. Less obvious, however, and less clear are the effects of references to sex by name or related pronoun, or of references to roles and occupations. There is general agreement that neutral item wording is desirable but also some doubt as to the net effect of such an effort--would making these changes affect test performance sufficiently to be worthwhile? (Datta, 1977).

Boldt (1983), in an outline study of research performed on various postsecondary assessment instruments, determined that aesthetic-philosophical, human relations, or female-oriented content tends to favor females, while practical affairs, science, or male-oriented content favors males. Donlon et. al. (1977), reviewing performance in grades 2, 5, 8, and 10 on selected major achievement tests, found that females performed less well on items with

greater numbers of references. Rowell and Hennen (1978) found that third-grade students taking a specially designed individual reading test did better on gender-specific rather than gender-neutral tests, but that neither boys nor girls did significantly better on tests containing items gender-specific to them.

Work done by Strassberg-Rosenberg and Donlon (1975) on results of the 1974 SAT indicated that reference to male or female characters and pronoun usage did not seem to affect the delta value or bias index of verbal items on that test. In math, there did not seem to be a direct relationship between pronoun usage and direction of bias for any particular item. Finally, there was no apparent relationship between stereotypical content and statistical bias. This study did, however, sustain the findings of earlier studies that items biased toward males or females tend to contain content representative of traditional male or female activities and interests, respectively.

Finally, a more recent study by Wedman and Stage (1983) further clouds the issue in reporting that, though sex differences in identifying items as being male or female tend to remain stable across male and female groups, it is difficult to identify items that give rise to sex differences in performance. Results of the study, based on the administration of specially designed items from the Swedish Scholastic Aptitude Test, did not support the hypothesis that it is possible to predict which items give rise to sex differences except in areas traditionally considered male or female. The authors of the study point out that it would be tempting to simply balance such items against each other in constructing a test if this would equalize effects on males and females, but that this approach would mean an acceptance, or even confirmation, of the traditional concepts of male and female roles.

Design

The present study was designed to improve upon previous research efforts. Because some previous researchers of this topic used existing tests and, often, data from convenience samples of examinees, they were hampered in their ability to disentangle item content and examinee characteristics from the specific effects of item gender. This study used specially constructed parallel versions of experimental items and the administration of those items to randomly equivalent groups in both single-gender and mixed-gender contexts. By exercising experimental control over both item and examinee characteristics, the design of this study enhanced isolation of the effects of gender-related item manipulation. For convenience and clarity, the male/female classification of examinees is referred to as examinee sex while the neuter/male/female classification of items is referred to as item gender throughout this paper.

Experimental sets of items (tryout units) were constructed in the content areas of English language usage and mathematics. Each of these experimental units was administered to approximately 250 examinees in randomly equivalent samples such that only one version of a particular item was administered to any one examinee.

Nine English language usage units were prepared. These were of two kinds: intact passages and discrete sentences. In the intact-passage format, items are embedded within a passage. The passage, a continuous piece of discourse, has underlined and numbered portions presented in one column. In a matching column, alternative responses are provided for each underlined portion from which the examinee must choose the one alternative that best expresses the idea, makes the statement appropriate for standard written English, or is worded most consistently with the style and tone of the passage

as a whole. Three versions of the experimental passage were created: a neuter version, a male version, and a female version. These were referred to as the gender versions of the item, or simply the item genders, to avoid confusion with another important variable, examinee sex. Editing changes to achieve the particular item gender orientation were made in references to possessions (e.g., articles of clothing) and in pronouns, rather than to the basic narrative, which concerned efforts to lose weight and was presented in a relatively gender-free manner. Twenty-eight test items were generated from the passage of 325 words. Of the 28 items, 25 were identical in all three gender versions, two had gender-related wording changes, and one asked the examinee to identify the gender of the passage.

For the discrete-sentence language units, twenty-one sentences were fashioned to assess various language situations involving sentence structure, grammar, punctuation, and diction. Each sentence was presented in three versions: neuter, male, and female. In some sentences, subject matter was neuter (i.e., described a situation, role, or occupation that could reasonably involve either male or female) as in the case, for example, of a teenager talking on the telephone or of a person playing a musical solo. In other cases, the subject matter dealt with a situation, role, or occupation most commonly perceived to involve females (as in cooking a holiday dinner) or males (as in playing poker). However, none of the discrete sentences required background knowledge that was expected to be differentially advantageous to either males or females in correctly answering the question. Editing changes thus consisted merely of changes in the gender of names, related pronouns, possessions, roles, or occupations in each of the three versions of the test questions.

The discrete-sentence items were then assembled into six tryout units. One unit presented the neuter versions of all twenty-one test items; one presented all the male versions; one presented all the female versions; and three presented a randomly ordered mix of all three versions with no two versions of any item appearing in any single unit. These mixed units were composed solely of gender-study items. Consequently, any one version of a discrete language item (neuter, male, or female) appeared twice among the six units, once in a single gender unit and once in a mixed unit.

Sixteen mathematics problems received the same kind of editorial manipulation as did the discrete-sentence language tryout units. All of the mathematics items tested the examinees' arithmetic and algebraic reasoning abilities by presenting practical problems in narrative form. As with the language items, single-gender units were constructed (all neuter, all male, or all female items). Experimental units contained 16 items, all of a word-problem type. Each of these units was administered twice. In addition, each item was embedded in a unit with a mixture of item types (not all word problems). For these mixed units, a male item version occupied position 3; a female version, position 6; and a neuter version, position 9, of a 16-item unit. The remaining items were not associated with the gender study. Thus each version of a math item (neuter, male, or female) appeared three times; twice in a single-gender unit, and once in a mixed unit.

The experimental items constructed for this study were written and reviewed by the researchers. These items were not subjected to the full review process required for ACT items that are used on actual tests and were not, therefore, as consistent in quality as regular ACT items. The format of the Mathematics and English passage-related items was the same as is normally used for the ACT Assessment program. Discrete English items were used only for this study; they do not appear in the ACT Assessment.

The experimental units, both single-gender and mixed, were administered to approximately 300 examinees (in randomly equivalent samples such that only one version was administered to any one examinee.

Analysis

Separate analysis was conducted of examinee responses to each item in each context. The loglinear analysis procedure employed is described in Appendix A and Table 1. It used four factors: item gender, examinee sex, examinee ability, and response. Responses were either correct or incorrect for a specific item. Examinees were divided into high-, average-, and low-ability groups based on their scores on the appropriate ACT subtest. Examinee sex was male or female; item gender was neuter, male or female. Differences in the difficulty of each item were examined to detect four types of effects: (1) those based on item gender differences; (2) those based on interactions between item gender and examinee sex, (3) those based on an interaction between the item gender and examinee ability; and (4) those based on interactions among item gender, examinee sex, and examinee ability.

Insert Table 1 about here

Only significant results for the second and fourth types of effects would indicate bias on the basis of examinee sex. That is because these analyses focus on the interaction of examinee sex with item difficulty and discrimination, respectively. The first and third effects do not involve examinee sex. Rather, they look at difficulty and discrimination differences, respectively, associated with the item gender.

The table-wide alpha level was set to approximately .05 for each unit in each context.

Results

Math Items

No math item showed significant differences in either difficulty or difficulty by ability (discrimination) on the basis of gender or of gender-by-sex interactions when the items were administered in single-gender units. Only one item showed such a difference and then only when it was placed in a mixed unit. That item, #60, had a difference in difficulty by gender with p-values of .59, .74 and .62 for neuter, male, and female versions, respectively. Item #60 is reproduced in all three versions in Appendix B. Its response patterns are provided in Table 2.

Insert Table 2 about here

Item #60 requires the examinee to use information provided to calculate the average rate of speed for a cyclist and then to use that information to determine the length of time a second cyclist travelling at the same rate of speed would require to travel a given distance. "A cyclist" and "a second cyclist" in the neuter version were replaced by "Sam" and "David" in the male version and "Jean" and "Paula" in the female version. Unfortunately, the version with female names had a typographic error which resulted in the inclusion of the name "Sam" in place of "Jean." Despite the confusion

obviously caused by this, the item was easier than the neuter version. It seems that for this item, the versions with names are easier.

Only this one item (of 16) showed any differences (based on testing 4 effects in the model) in this one context (of two). Given the table-wide alpha-level of .032, one would expect that about 2 tests in each 16 x 4 table (64 tests) would be statistically significant by chance alone. This single significant result is well within that expectation.

Nevertheless, as a check, the two administrations of each single-gender unit in math (which had been combined in the previous analysis) were compared to one another. Since both the items and contexts were identical and examinees were randomly assigned, the only cause of differences should be chance. This appeared to be the case with only one item, #55, showing a single significant difference. Item #55 is also reproduced in Appendix B.

English Discrete Items

No English discrete item showed significant differences in discrimination on the basis of gender or gender-by-sex interactions or in difficulty on the basis of gender-by-sex interactions whether placed in a single-gender or mixed-gender unit. Several items, however, showed differences in difficulty by gender. There were 4 such items (#30, #31, #33, and #37) in single-gender units, and 5 items (#32, #33, #36, #40 and #48) in mixed-gender units. Table 3 provides the likelihood ratio chi-squares for these items. The text of the items is provided in Appendix C.

Insert Table 2 about here

Item numbers served as identification in this study. Therefore the comparison of item numbers across the two sets indicates that only item #33 showed these differences in both contexts. However, because item numbers were used to identify the multiple versions of items across contexts, they did not reflect the ordinal position of the item within a particular context. Table 4 provides p-values and the ordinal location of items that had significant differences. Each English discrete tryout unit had 21 items. There did not appear to be any consistent relationship between ordinal position and item difficulty.

Insert Table 3 about here

Examination of item text suggested that the reasons that individual items showed differential difficulty probably varied. Item #31, for example, was a flawed item: it had two correct answers. Both the keyed answer, NO CHANGE, and option H, "think" are defensible. Recalculating the p-values using either answer as correct results in p-values of .89, .98, and .95 for the neuter, male, and female versions, respectively. The original analysis may have detected some differential attractiveness of the two correct answers based on the specific gender version.

Item #30 focused on the phrase "incorporated... into". As with item #31, the problem appeared to be the differential attractiveness of a second plausible foil "incorporated... with". Because of the lack of a passage context, examinees may have interpreted the item to refer to including something with, as opposed to within, films. They may have missed the comprehension of a "plot line" as something that is not an independent object and cannot, therefore, exist other than within a communication vehicle, in

this case a film. This misconception appears to have been most attractive for the neutral version, which refers to "a group of French film directors"; somewhat less attractive in the male version, "Another Frenchman"; and least attractive for the female version , "another French woman". It is tempting to speculate that students ignored the implications of "another" and interpreted the male version with reference to the more generic "Frenchmen."

Item #36 tests the past tense of the verb "to begin", i.e. began. The male version, "he began" was easier (p = .91) than the female version "she began" (p = .86) or the neuter "they began (p = .84). A similar pattern occurred for item #40 which tested the who/whom distinction together with the past tense of the verb to fight. Again the male version, "the worst of the kings who fought", (p = .70) was easier than the female version, "the worst of the queens who fought", (p = .57) or the neutral version, "the worst of those who fought" (p = .53). The most attractive foil in each case was "whom fought". However, the third most popular foil, "who have been fighting", selected by 4% in the male version, 10% in the female version, and 13% in the neuter version was actually a defensible second key. It may have been that students think about kings who fought, historically, and queens who have been (and perhaps still are) fighting. In any case, adding the responses to the two foils resulted in p-values of .66 for the neuter, .67 for the female, and .74 for the male versions, much smaller differences than when only the originally keyed response was considered. Interestingly, this differential preference for the two correct responses did not occur when the items were presented in single-gender units (see Table 5).

Insert Table 5 about here

As with items #30 and #31, item #37 had a single powerful distracter. In this case, the original wording, "operates in cooking just as it does in the composition of a poem..." was correct. Substitution of the attractive foil, "in a manner similar to that", would have left "operates in cooking in a manner similar to that in the composition of a poem", an awkward construction at best. Where the item was presented in the neutral version, it had a p-value of .73, very similar to that of the female version (p = .70) which simply added the introductory phrase "For Janet, ". The male version, which added the introductory phrase "For George, " appeared considerably more difficult, perhaps due to some vestigial cognitive dissonance associated with men cooking.

Items #32 and #33 were in the same sentence: "The rules in bridge call for formal rounds of bidding, but any other attempt to communicate information concerning the cards a player holds are forbidden." The first underlining is correct as it stands; the second should be replaced by "is". Item #32 did not show gender differences in the single-gender units, but did in the mixed-gender units. In this case, the neutral version cited above was the easiest (p = .63). The next most attractive foil, "to try and", was selected by 26% of the examinees, and the third most attractive, "you make to" was selected by 10%.

In the female version, " a player", was replaced by "Deb and Sue" and the percentages of examinees selecting the key and top two foils were 52%, 27% and 13%, respectively, with the fourth foil increasing from 1% in the neutral to 8% in the female version. The male version substituted the names "Ralph and Rich". Percentages selecting the key and top two foils were 43%, 28%, and 20%, respectively, with the second most attractive foil, "you make to",

discriminating negatively. It appears that this foil may be a defensible second key, particularly given the substitution of the two names for the generic " a player," and that examinees were more sensitive to it when the names were male than when they were female.

Item #33 was the only item to show gender differences in both single-gender and mixed tryout units. Proportions of examinees selecting each foil for the three versions in two contexts are provided in Table 6.

Insert Table 6 about here

These appear to be differences not only among the gender versions (neutral easiest) but also between the single-gender and mixed-gender contexts. One may speculate that the generic "a player" was seen as more consistent with the eternal "is" while the use of specific names suggested the possibility that the rules were also situation-specific and therefore, communications "have been" forbidden, at least for the moment. While this interpretation would be strange in the context of a correct response to the previous item:

> "any other attempt to communicate information concerning the cards Deb and Sue hold have been forbidden",

in the context of the alternate foil, the sentence becomes complex enough that weaker students may have been unable to determine its key structural features:

> "any other attempt you make to communicate information concerning the cards Deb and Sue hold have been forbidden"

In any case, the replacement of a single "player" with two individuals "Deb and Sue" in between subject (attempt) and verb (is forbidden) clearly makes it more difficult for the examinee to correctly establish subject-verb agreement.

Differences between the single-gender and mixed-gender contexts for this item are also perplexing: the neuter and female versions were easier in the single-gender context, but the male version was easier in the mixed-gender context. These differences appear to be related to the relative attractiveness of the NO CHANGE and "have been" options in the neuter and female versions.

Items #48 and #49 were also in a single sentence. Item #49 did not show any differences, but item #48 was more difficult in the male version in the mixed-gender content only. Although the male and female versions of the item are a bit longer than the neutral one ("teenage... a friend" replaced by "son/daughter... a school locker partner"), it is unclear why the male version (son) appeared more difficult than the female one (daughter).

English Passage-based Items

No English passage-based item showed a significant difference in difficulty or discrimination on the basis of a gender-by-sex interaction. Three items (#9, #20, and #28) showed differences in difficulty and one (#19) showed a difference in discrimination based on gender alone. Tables 7 and 8 provide the likelihood ratio chi-square and the p-values, respectively, for these items. The passage-based items had fixed locations in the passage; their item numbers reflect their ordinal position.

Insert Tables 4 and 5 about here

Item #9 tested subject-verb agreement. In all 3 versions, the grammatical structure was the same: nobody calls to say that they has.... The keyed responses varied: "he or she has" (neutral), "he has" (male) and "she has" (female). Since the sex of "nobody" was indeterminate, students found the neuter version, "he or she", the easiest (p = .81). The generic implications of "he" (p = .64) appeared to make that choice more readily acceptable than the specific "she" (p =.31), even though in each case the keyed response was the only grammatically acceptable one.

Items #19 and #20 appeared in the same sentence (neuter version):

that my cheeks are more fulfilled ... and
19

that fitting into my pants seem ...
20

Item #19 was the only item in the study to show significant discrimination differences based on item gender. This is not to be confused with "bias" as these differences did not involve examinee sex. The loglinear approach uses the ability-by-response interaction term to index discrimination. Thus, the significant effect for this item indicates differential response by ability on the basis of item gender. This is illustrated in Figure 1. The text for the item varied for the three versions:

Type	Text	Key
Neuter	my cheeks are more fulfilled	(are fuller)
Male	my shoulders are more broadened	(are broader)
Female	my face is more fulfilled	(is fuller)

The male version ("broader", as opposed to "fuller") appeared easier overall, especially for the lower ability group. For reasons best known to themselves, the average group found the "NO CHANGE" alternative unusually attractive for the male version of the item; fifteen percent of them selected it (as opposed to 6% for the neuter version and 7% for the female version). Although the mechanism causing this response pattern is uncertain, it seems clear that changes to the item introduced multiple and interactive types of differences: [cheeks/shoulders/face] [are/is] [fulfilled, fuller/broaden, broader]. In retrospect, it is not surprising that the items functioned differently.

Item #20 needed to have "fitting into" deleted for the "seem" to become grammatically correct ("that fitting into my pants seem"). "Pants" in the neuter version was replaced with "trousers" in the male version and "skirts" in the female version. For some reason, the construction with trousers ("my trousers seem") was easier ($p = .45$) than the same construction either with pants ($p = .29$) or with skirts ($p = .27$).

Item #28 asked the examinee to identify the age/gender status of the individual referred to as "I" in the passage. The intent of the item was to determine whether in fact the intended treatment (gender differentiation of the passage) had been detected. Examinees had no difficulty determining that "I" referred to an adult rather than a child in any of the versions, nor did they have difficulty identifying the female version ($p = .88$).

The male and neuter versions appeared more difficult to identify. Forty-six percent of all examinees correctly identified the male version, 19% thought it was female, and 32% thought it was neuter. The neuter version was also correctly identified by 46% of all examinees, while 16% thought it was male and 33% thought it was neuter.

Discussion

Superficial gender-related item modifications such as those that were made for this study do not appear to create sex bias. In fact, reference to male or female names, pronouns, possessions, or occupations in the place of neutral language had no demonstrable effect at all on examinee performance on mathematics word problems.

For English items, the gender-related modifications affected item difficulty significantly in about 17% (12 of 70) of the instances in which it was examined. In each case, however, the difficulty difference was the same for male and female examinees; sex bias was not introduced.

Examination of the statistics for items that showed shifts in difficulty did not suggest that items of a particular gender category are consistently harder or easier than items in other categories. Where the text of the item as presented in the three gender versions suggests an explanation of the examinee performance pattern (which is not always the case), that explanation usually relates to a real change in the difficulty of the item itself. For example, the change may have inadvertently made the sentence structure more complex. Such a change may occasionally result in differential changes in the difficulty of the item for high, average, and lower-performing examinees (i.e., a difference in discrimination). Usually, however, the difficulty

change did not appear to interact with ability level. English language item difficulty appears to relate to examinee familiarity with particular syntactic structures, idiomatic expressions, or other linguistic features. To the extent that the use of gender-neutral language increases the complexity of the item, for example, it may tend to make the item more difficult for examinees.

In creating test items, test developers must remain aware of many, often conflicting goals. Coverage of content areas and cognitive levels vies with the needs to maintain appropriate distributions of item difficulty and discrimination levels and to guard against overlapping content and cuing. Over this collage of concerns, the test developer must superimpose considerations of social and ethical values. Test items must not be offensive or upsetting to any group. They must provide fair representation of the social context, avoiding stereotypes, yet not venture so far beyond the norm as to hinder the examinee's attention to the essential features of the test question.

It is, perhaps, comforting to know that while it remains appropriate to balance references to males and females in tests for a variety of social and ethical reasons, failure to do so is not likely to hamper the performance of examinees of either sex on a differential basis. Neither does it seem necessary to couch all test items in neutral terms, a task that often turns out to be far more difficult than it may appear and that may in some cases hamper the content validity of the item.

To the extent that the focus of the item is neutral and that care is taken to retain the intended content focus and level of linguistic complexity of English language items, male, female, and neutral wording all appear appropriate vehicles for conveying the test question and eliciting the examinee's response.

References

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press.

Boldt, R. F. (1983). Status of research on item content and differential performance on tests used in higher education. Educational Testing Service: Princeton, NJ.

Carlson, J. E. (1986). Detection of item parameter shifts between two time points; one method and some problems. Unpublished manuscript.

Carlson, J. E. & Spray, J. A (1986). Analysis of contingency tables involving multiple-response data. Unpublished manuscript.

Datta, L. E. (1977). he and she: Sex fairness in selection and guidance based on educational testing. ED 143682.

Donlon, T. F. and Others (1977). Performance consequences of sex bias in the content of major achievement batteries. Final Report. Educational Testing Service: Princeton, NJ.

Feinberg, S. E. (1977). The analysis of cross-classified categorical data. Cambridge, MA: MIT Press.

Kok, F. G., Mellenbergh, G. J., & Van Der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logist method. Journal of Educational Measurement, 22, 295-303.

McCarthy, K. (1975). Sex bias in tests of mathematics. Doctoral dissertation, City University of New York.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Rowell, E. H. & Hennen, F. G. (1978). The interaction between sex referents in test items and reading performance on individualized reading tests. ED 191009.

Searle, S. R. (1971). Linear models. New York: Wiley

Spray, J. A. & Carlson, J. (1986, April). Comparison of loglinear and logistic regression models for detecting changes in proportions. Paper presented to the meeting of the American Educational Research Association, San Francisco.

Strassberg-Rosenberg, B. & Donlon, T. F. (1975). Content influences on sex differences in performance on aptitude tests. National Council on Measurement in Education.

Van Der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wijn, N. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

Wedman, I. & Stage, C. (1983). The significance of contents for sex differences in test results. Scandanavian Journal of Educational Research, 27, 49-71.

Table 1

Gender Study Full Model[1]

Full Model:

Code	Effect	Content
S	Examinee sex	Proportion who are male, female.
G	Item gender	Proportion who are neuter, male, female.
A	Ability[2]	Proportion who are high, average, low.
R	Response Category responses.	Proportion who give correct, incorrect
SG	Examinee sex by item gender	Proportion of males, females who responded to each item gender.
SA	Examinee sex by ability	Proportion of males, females in each ability group.
SR	Examine sex by response	proportion of males, females who responded correctly.
GA	Item gender by ability	Proportion of items of each gender category responded to by each ability group.
GR	Item gender by response	Proportion of items of each gender category responded to correctly.
AR	Ability by response	Proportion of correct responses for each ability level.
SGA	Sex by gender by ability	Proportion of males, females in each ability group responding to items of each gender type.
SGR	Sex by gender by response	Proportion of males, females responding correctly to items of each gender type.
SAR	sex by ability by response	Proportion of males, females in each ability group responding correctly.
GAR	Gender by ability by response	Proportion of each ability group responding correctly in each item gender category.
SGAR	Sex by gender by ability by response	Proportion of males, females of each ability level responding correctly to items of each gender type.

[1] This complete analysis was applied to each item in each context.

[2] Ability groups were formed separately for males and females such that the examinees whose scores were in the top and bottom 27% of each sex formed the high and low groups, respectively, while the middle 46% formed the middle group.

Table 2
Item Response Patterns for Mathematics Item #60
in Nine Settings

	Single Gender						Single Gender						Mixed					
Number 60 - Neutral																		
	11	1	2*	3	4	5	0	11	1	2*	3	4	5	0	9	1	2*	3	4	5
UPP	2	34	10	0	3	1	UPP	2	79	14	2	1	2	UPP	2	76	13	3	5	
MID	2	60	25	7	3	2	MID	7	55	28	5	3	1	MID	6	53	31	6	5	
LOW	11	40	30	11	6	2	LOW	13	43	30	11	2	2	LOW	2	40	40	9	9	
TOT	4	65	20	5	4	2	TOT	6	62	23	5	2	2	TOT	4	59	26	5	6	
Number 60 - Male																				
	11	1	2*	3	4	5	0	11	1	2*	3	4	5	0	3	1	2*	3	4	5
UPP	0	91	7	0	0	2	UPP	0	96	3	0	1	0	UPP	1	93	3	2	0	
MID	3	69	26	3	0	1	MID	3	58	28	3	5	2	MID	2	73	20	3	2	
LOW	6	43	37	10	4	0	LOW	13	31	46	3	8	0	LOW	13	39	38	6	6	
TOT	3	72	22	2	1	1	TOT	4	69	21	2	4	1	TOT	3	74	18	3	2	
Number 60 - Female																				
	11	1	2*	3	4	5	0	11	1	2*	3	4	5	0	6	1	2*	3	4	5
UPP	1	85	6	3	2	2	UPP	0	83	10	5	2	0	UPP	1	77	16	0	2	
MID	4	68	21	4	2	1	MID	2	59	32	6	0	2	MID	4	66	26	7	4	
LOW	11	34	32	16	5	0	LOW	9	44	35	9	4	0	LOW	15	50	24	7	4	
TOT	4	68	18	6	2	1	TOT	3	64	25	6	2	1	TOT	5	62	23	4	3	

Table 3
Likelihood Ratio Chi-squares for
English Discrete Items in
Single-Gender and Mixed-Gender Units
(N = 816 and 770, respectively)

	Difficulty[1]		Discrimination[2]	
	Gender	Gender X Sex	Gender	Gender X Sex
Degrees of Freedom	2	2	4	4
Single Gender Units				
Item #				
29.	-	-	-	-
30.	21.83	-	-	-
31.	30.13	-	-	-
32.	-	-	-	-
33.	23.92	-	-	-
34.	-	-	-	-
35.	-	-	-	-
36.	-	-	-	-
37.	17.23	-	-	-
38.	-	-	-	-
39.	-	-	-	-
40.	-	-	-	-
41.	-	-	-	-
42.	-	-	-	-
43.	-	-	-	-
44.	-	-	-	-
45.	-	-	-	-
46.	-	-	-	-
47.	-	-	-	-
48.	-	-	-	-
49.	-	-	-	-
Mixed Gender Units				
29.	-	-	-	-
30.	-	-	-	-
31.	-	-	-	-
32.	26.84	-	-	-
33.	17.72	-	-	-
34.		-	-	-
35.		-	-	-
36.	18.71	-	-	-
37.	-	-	-	-
38.	-	-	-	-
39.	-	-	-	-
40.	23.01	-	-	-
41.	-	-	-	-
42.	-	-	-	-
43.	-	-	-	-
44.	-	-	-	-
45.	-	-	-	-
46.	-	-	-	-
47.	-	-	-	-
48.	21.48	-	-	-
49.	-	-	-	-

[1] Only those chi-squares exceeding a critical value of 15.20 ($\alpha = .0005$ with 2, = degrees of freedom) are printed.

[2] Only those chi-squares exceeding a critical value of 20.00 ($\alpha = .0005$ with 4, = degrees of freedom) are printed.

Table 4
P-Values and Ordinal Position for English
Discrete Items Showing
Difficulty Differences by Gender

Item #	In Single Gender Units		In Mixed Units	
	p-value	Ordinal Position N,M,F[1]	p-value	Ordinal Position N,M,F[1]
29.	-	-	-	-
30.	55,73,66	2,16,13	-	-
31.	79,69,75	3,17,8	-	
32.	-	-	61,43,52	3,18,15
33.	75,56,62	5,19,10	70,63,55	14,9,16
34.	-	-	-	-
35.	-	-	-	-
36.	-	-	84,91,86	4,18,7
37.	73,57,70	9,3,18	-	-
38.	-	-	-	-
39.	-	-	-	-
40.	-	-	53,70,57	10,5,6
41.	-	-	-	-
42.	-	-	-	-
43.	-	-	-	-
44.	-	-	-	-
45.	-	-	-	-
46.	-	-	-	-
47.	-	-	-	-
48.	-	-	76,60,74	3,1,2
49.	-	-	-	-

[1] Neuter, Male, Female

Table 9
Item Response Patterns for English Language Discrete
Item #40 in Six Settings

Single Gender

Number 40 - Neutral

12	1*	2	3	4	5	0
UPP	82	4	12	1	0	0
MID	60	11	25	4	0	0
LOW	32	20	41	7	0	0
TOT	61	10	24	4	0	0

Number 40 - Male

5	1*	2	3	4	5	0
UPP	79	8	13	0	0	0
MID	62	8	27	4	0	0
LOW	39	10	47	4	0	0
TOT	62	8	27	3	0	0

Number 40 - Female

20	1*	2	3	4	5	0
UPP	77	7	13	3	0	0
MID	71	6	20	3	0	0
LOW	34	16	46	4	0	0
TOT	65	8	24	3	0	0

Mixed

10	1*	2	3	4	5	0
UPP	80	8	10	2	0	0
MID	44	16	36	4	0	0
LOW	27	13	42	17	0	0
TOT	53	13	29	5	0	0

5	1*	2	3	4	5	0
UPP	87	1	9	3	0	0
MID	69	4	24	4	0	0
LOW	50	10	32	8	0	0
TOT	70	4	21	5	0	0

6	1*	2	3	4	5	0
UPP	71	7	16	5	0	0
MID	57	10	27	6	0	0
LOW	33	13	47	7	0	0
TOT	57	10	27	6	0	0

Table 6
Item Response Patterns for English Language Discrete
Item #33 in Six Settings

	Single Gender						Mixed					
Number 33 - Neutral												
	5	1	2	3	4*	5	0	14	1	2	3	4*	5	0
UPP	7	0	1	91	0	0		UPP	11	0	0	89	0	0
MID	15	2	5	78	0	0		MID	27	1	3	70	0	0
LOW	27	20	15	39	0	0		LOW	34	6	12	46	2	0
TOT	15	4	6	75	0	0		TOT	24	2	4	70	0	0
Number 33 - Male														
	19	1	2	3	4*	5	0	9	1	2	3	4*	5	0
UPP	14	0	3	83	0	0		UPP	17	0	1	81	0	0
MID	32	3	16	48	1	0		MID	24	4	13	59	1	0
LOW	18	16	22	41	0	2		LOW	24	7	24	44	0	0
TOT	25	5	14	56	0	0		TOT	22	3	11	63	0	0
Number 33 - Female														
	10	1	2	3	4*	5	0	16	1	2	3	4*	5	0
UPP	17	0	11	72	0	0		UPP	16	1	7	76	0	0
MID	18	2	13	65	1	0		MID	26	3	18	53	0	1
LOW	23	9	25	43	0	0		LOW	35	17	25	22	0	0
TOT	19	3	15	62	1	0		TOT	24	5	16	55	0	0

Table 7
Likelihood Ratio Chi-square for
English Passage Items (N = 742)

	Difficulty[1]		Discrimination[2]	
	Gender	Gender X Sex	Gender	Gender X Sex
Degrees of Freedom	2	2	4	4
Item #				
1	-	-	-	-
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-
5	-	-	-	-
6	-	-	-	-
7	-	-	-	-
8	-	-	-	-
9	148.46	-	-	-
10	-	-	-	-
11.	-	-	-	-
12.	-	-	-	-
13.	-	-	-	-
14.	-	-	-	-
15.	-	-	-	-
16.	-	-	-	-
17.	-	-	-	-
18.	-	-	-	-
19.	-	-	22.91	-
20.	19.96	-	-	-
21.	-	-	-	-
22.	-	-	-	-
23.	-	-	-	-
24.	-	-	-	-
25.	-	-	-	-
26.	-	-	-	-
27.	-	-	-	-
28.	126.59	-	-	-

[1] Only those chi-squares exceeding a critical value of 15.20 (α= .0005 with 2, = degrees of freedom) are printed

[2] Only those chi-squares exceeding a critical value of 20.00 (α= .0005 with 4, = degrees of freedom) are printed.

Table 8
P-Values and Biserial Correlations
for English Passage Items Showing
Significant Difficulty or Discrimination
Differences

Item #	Difficulty[1]		Discrimination[2]	
	Gender	Gender X Sex	Gender	Gender x Sex
	N,H,F	H: N,H,F F: N,H,F	N,H,F	H: N,H,F F: N,H,F
1.	-	-	-	
2.	-	-	-	-
3.	-	-	-	-
4.	-	-	-	
5.	-	-	-	-
6.	-	-	-	-
7.	-	-	-	-
8.	-	-	-	
9.	81,64,31	-	-	-
10.	-	-	-	-
11.	-	-	-	-
12.	-	-	-	
13.	-	-	-	-
14.	-	-	-	-
15.	-	-	-	-
16.	-	-	-	-
17.	-	-	-	-
18.	-	-	-	-
19.	-	-	37,30,61	
20.	29,45,27	-	-	-
21.	-	-	-	-
22.	-	-	-	-
23.	-	-	-	-
24.	-	-	-	-
25.	-	-	-	-
26.	-	-	-	-
27.	-	-	-	-
28.	48,46,88	-	-	-

[1] p-value
[2] biserial correlation

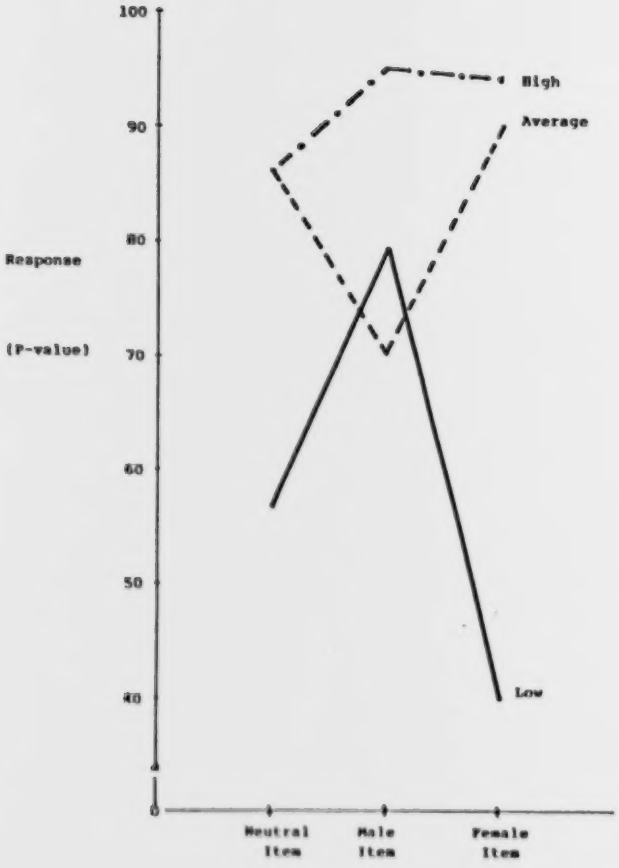

Figure 1

Interaction of Item Gender
with Ability Group and Response
For Item #19 (English Passage-related)

Appendix A:

Loglinear Methodology

The following discussion is adapted from the work of Carlson (1986), Carlson & Spray (1986), and Spray & Carlson (1986).

Within the context of the present problem, if each item in an n-item test is treated as a separate entity, item responses can be categorized into n 4-way contingency tables according to four categorical variables: examinee sex (S) with two levels, male and female; item gender (G) with three levels, neuter, male, and female; item response (R) with two levels, correct and incorrect; and ability (A) with 3 levels, low, average, and high (see Table 1). The frequencies within each cell of a table can be analyzed using a technique called loglinear (LL) modeling or analysis (Bishop, Fienberg & Holland, 1985; Fienberg, 1977). The inclusion of the ability variable in the table allows a test of proportion-correct differences across examinee sex and item gender for given levels of ability. The detection of these differences has been termed a test of "nonuniform bias" or "conditional methods of bias detection" by Mellenbergh (1982), Van Der Flier et al. (1984) and Kok et al. (1985). For this analysis, ability groups were formed based on an external criterion, the ACT score in the appropriate subject area (English or Mathematics). Groups were formed separately for male and females so that approximately 27% of each fell into the top and bottom (high and low) groups and the middle 46% into the average group.

The LL procedure consists of modeling the expected value of the log of the cell frequencies or proportions in terms of main, or marginal, effects and

interaction effects. The models have an "ANOVA-like" appearance to them except that there is no random error term in the model since there is only one observed frequency per cell, the number of individual examinees who fall into that category. Hence, there is no "within-cell" variation.

One apparent difference between ANOVA and LL models is a function of the way in which these models are typically described in the statistical literature. The LL analysis of cell frequencies is usually conceptualized in terms of comparisons of goodness-of-fit indices of a series of hierarchical models (with and without various interaction terms present). Although ANOVA can be conceptualized in an analogous fashion (e.g., Searle, 1971), the description in applied statistics texts is more often in terms of partitions of variation based on a single model. Each model for the frequency data yields expected counts for each cell and the goodness-of-fit statistic compares the expected counts with the observed counts. The most commonly used statistics for this comparison are the Pearson chi-square statistic and the likelihood-ratio chi-square statistic (abbreviated as G^2). The latter is preferred because differences between values of this statistic for hierarchical models are themselves distributed as chi-square random variables (Bishop, Fienberg, & Holland, 1975).

Another difference between ANOVA and LL models is that where the parameters of the ANOVA model are naturally additive, those for frequencies from a multi-way contingency table are multiplicative. Hence, logarithms of the multiplicative parameters are taken to get an additive, linear model. A cell mean in an ANOVA model, for example, is the sum of parameters associated with each main effect and interaction. An expected cell frequency under a model that assumes independence in a multi-way contingency table, on the other hand, is a product of parameters associated with each main effect.

Recalling the product rule for probabilities of the occurrence of independent events, it can be seen that if sex, gender, ability and response are assumed to be independent,

$$P_{ijkl} = P_{i\cdots}P_{\cdot j\cdot\cdot}P_{\cdot\cdot k\cdot}P_{\cdot\cdot\cdot l} \text{ where } i = 1,2;\ j = 1,2,3;\ k = 1,2,3;\ l = 1,2.$$

In words, the probability of a randomly selected observation being in any cell ijkl is the product of the four probabilities. The expected cell count under this model would be

$$m_{ijkl} = Np_{ijkl} = Np_{i\cdots}p_{\cdot j\cdot\cdot}p_{\cdot\cdot k\cdot}p_{\cdot\cdot\cdot l},$$

which illustrates the multiplicative relationship mentioned previously. An alternative formulation of m_{ijkl} is

$$m_{ijkl} = N\frac{m_{i\cdots}}{N}\frac{m_{\cdot j\cdot\cdot}}{N}\frac{m_{\cdot\cdot k\cdot}}{N}\frac{m_{\cdot\cdot\cdot l}}{N} = \frac{m_{i\cdots}m_{\cdot j\cdot\cdot}m_{\cdot\cdot k\cdot}m_{\cdot\cdot\cdot l}}{N^3}$$

where

$$m_{i\cdots} = \sum_j\sum_k\sum_l m_{ijkl} = Np_{i\cdots},$$

$$m_{\cdot j\cdot\cdot} = \sum_i\sum_k\sum_l m_{ijkl} = Np_{\cdot j\cdot\cdot},$$

$$m_{\cdot\cdot k\cdot} = \sum_i\sum_j\sum_l m_{ijkl} = Np_{\cdot\cdot k\cdot},$$

and $$m_{\cdot\cdot\cdot l} = \sum_i\sum_j\sum_k m_{ijkl} = Np_{\cdot\cdot\cdot l}.$$

Taking logarithms and letting ℓ_{ijkl} represent the log of the expected cell count gives the loglinear model,

$$\ell_{ijkl} = \ln(m_{ijkl}) = \ln \frac{m_{i+++} m_{+j++} m_{++k+} m_{+++l}}{N^3}$$

$$= -\ln(N^3) + \ln(m_{i+++}) + \ln(m_{+j++}) + \ln(m_{++k+}) + \ln(m_{+++l}) .$$

Recall that this model assumes independence of all four categorical variables. The four terms in the model can be thought of as representing an overall effect and main effects of the three categorical variables and is usually written as

$$\ell_{ijkl} = \mu + \mu_{S(i)} + \mu_{G(j)} + \mu_{A(k)} + \mu_{R(l)}. \qquad (1)$$

where $\mu = -\ln(N^3)$, $\mu_{S(i)} = \ln(m_{i+++})$, $\mu_{G(j)} = \ln(m_{+j++})$ $\mu_{A(k)} = \ln(m_{++k+})$ and $\mu_{R(l)} = \ln(m_{+++l})$. As in the analysis of variance, these parameters are not estimable without restrictions or constraints which are

$$\sum_i \mu_{S(i)} = \sum_j \mu_{G(j)} = \sum_k \mu_{A(k)} = \sum_l \mu_{R(l)} = 0.$$

If the assumption of independence of all four variables is removed, additional terms are added to the model in order to account for these dependencies. All LL models are special cases of a general loglinear model (Fienberg 1977) written as:

$$\mu_{ijkl} = \mu + \mu_{S(i)} + \mu_{G(j)} + \mu_{A(k)} + \mu_{R(l)} + \mu_{SG(ij)} + \mu_{SA(ik)} +$$

$$\mu_{SR(il)} + \mu_{GA(jk)} + \mu_{GR(jl)} + \mu_{AR(kl)} + \mu_{SGA(ijk)} + \mu_{SGR(ijl)} +$$

$$\mu_{SAR(ikl)} + \mu_{GAR(jkl)} + \mu_{SGAR(ijkl)}. \qquad (2)$$

Model (2) is often referred to as the saturated model because it contains exactly as many parameters to be estimated as there are cells in the table [i.e., [1 + (I - 1) + (J - 1) + (K - 1) + (L - 1) + (I - 1) (J - 1) + (I - 1) (K - 1) + (I - 1) (L - 1) + (J - 1) (K - 1) + (J - 1) (L - 1) + (K - 1) (L - 1) + (I - 1) (J - 1) (K - 1) + (I - 1) (J - 1) (L - 1) + (I - 1) (K - 1) (L - 1) + (J - 1) (K - 1) (L - 1) + (I - 1) (J - 1) (K - 1) (L - 1)] versus [IJK]] ; hence there are no degrees of freedom for tests of lack of fit. (Degrees of freedom = # cells - # parameters estimated).

By setting various terms in model (2) equal to zero, other models which fall in between the independent case of model (1) and the saturated case of model (2) are possible. For example, a simpler model than model (2) can be postulated by setting $\mu_{SGAR} = 0$ and calculating a goodness-of-fit statistic for the expected cell counts from the hypothesized model. If this statistic is significant at some given level, the conclusion is that the simpler model does not fit the data observed; the dependencies must involve all four variables in a significant, quadruple interaction term.

We denote the likelihood-ratio chi-square statistic for the test of $\mu_{SGAR} = 0$ by G^2_{SGAR}. If this statistic is nonsignificant, usual practice is to fit a model without the four-way interaction but including any three-way interaction terms of interest. In this way many other models are possible as

long as two "rules" are obeyed. First, the formulation of alternative models is limited to those included in "hierarchical sets in which higher-order terms may be included only if the related lower-order terms are included" (Fienberg, 1977, p. 39). An example of two models that would be included in the same hierarchical set would be models (3) and (4) below since

$$\ell_{ijkl} = \mu_S + \mu_G + \mu_A + \mu_R + \mu_{SA} + \mu_{RA} \qquad (3)$$

and

$$\ell_{ijkl} = \mu + \mu_S + \mu_G + \mu_A + \mu_R + \mu_{SA} + \mu_{RA} + \mu_{SR} \qquad (4)$$

follow this rule while, for example:

$$\ell_{ijkl} = \mu + \mu_S + \mu_R + \mu_{RA}$$

and $\ell_{ijkl} = \mu + \mu_A + \mu_{SR}$ do not.

The second requirement is that terms involving fixed marginal totals must be included in the model being tested. Within the context of the present problem, the marginals are fixed for item gender, examinee sex, and ability; therefore, the terms μ_G, μ_A, μ_S and μ_{GA}, μ_{SG}, μ_{SA}, and μ_{SGA} must be included.

As mentioned previously, a significance test of a postulated model is in the form of a goodness-of-fit test via a likelihood-ratio chi-square statistic, denoted by G^2. As each term is added to the model and the model is tested against the observed data, a G^2 statistic is obtained. The difference

between the G^2 statistic for this model and the previous model is then a chi-square statistic for a test of the significance of the added term. This procedure of successively fitting hierarchical models is more completely discussed by Bishop, Fienberg and Holland (1975) and Fienberg (1977). These sources also give methods for computing expected cell counts for each hypothesized model, as well as the constraints imposed.

For the present problem four pairs of models were compared. The fully saturated model (Model 2 above) was compared to a model in which the four-way interaction term, SGAR, has been omitted (Model 5). If the chi-square difference between these two models is significant, this suggests that the interaction among examinee sex, item gender, ability group, and correct responses is necessary to describe the item's functioning. That is, that the pattern of p-values for males in high, average and low ability groups on neuter, male, and female items is different than that for females under those conditions. In effect, this forms a test for discrimination differences (ability by response interaction) based on an examinee sex by item gender interaction.

If that comparison is not significant, Model 5 can be compared with a model which drops the GAR term (Model 6). This comparison, if significant, indicates that there are item gender differences in the response patterns for the three ability groups. That is, that discrimination varies with item gender.

$$\mu_{ijkl} = \mu_S + \mu_G + \mu_A + \mu_R + \mu_{SG} + \mu_{SA} + \mu_{SR} + \mu_{GR} + \mu_{GA} + \mu_{AR} + \mu_{SGA} + \mu_{SGR} + \mu_{SAR} + \mu_{GAR} \qquad (5)$$

$$\ell_{ijkl} = \mu_S + \mu_G + \mu_A + \mu_R + \mu_{SG} + \mu_{SA} + \mu_{SR} + \mu_{GR} + \mu_{GA} + \mu_{AR} + \mu_{SGA} + \mu_{SGR} + \mu_{SAR} \quad (6)$$

Together, these tests identify significant differences associated with the interaction of item difficulty (response) with examinee ability. The first comparison considers this interaction in light of examinee sex by item gender interactions; the second collapses the examinee sex categories to look at item gender differences alone. The interaction of item difficulty with examinee ability is interpretable as an index of item discrimination.

Models 7, 8 and 9 provide a similar set of comparisons focusing on item difficulty.

$$\ell_{ijkl} = \mu_S + \mu_G + \mu_A + \mu_R + \mu_{SG} + \mu_{SA} + \mu_{SR} + \mu_{GR} + \mu_{GA} + \mu_{AR} + \mu_{SGA} + \mu_{SGR} \quad (7)$$

$$\ell_{ijkl} = \mu_S + \mu_G + \mu_A + \mu_R + \mu_{SG} + \mu_{SA} + \mu_{SR} + \mu_{GR} + \mu_{GA} + \mu_{AR} + \mu_{SGA} \quad (8)$$

$$\ell_{ijkl} = \mu_S + \mu_G + \mu_A + \mu_R + \mu_{SG} + \mu_{SA} + \mu_{SR} + \mu_{GA} + \mu_{AR} + \mu_{SGA} \quad (9)$$

The comparison of Models 7 and 8 drops the SGR term. If this difference is significant, it indicates a significant difficulty difference (response) in the examinee sex by item gender interaction. If this difference is not significant, a comparison of Models 8 and 9 tests the effect of item gender

alone on item difficulty (response). As with ANOVA, higher order tests need to be evaluated first since a significant higher-order interaction may affect the interpretation of a less complex effect.

Each of the four model comparisons discussed above (5 with 2, 6 with 5, 8 with 7, and 9 with 8) was made for each item as it appeared in each type of context (single-gender or mixed). Thus there were 32 analyses for mathematics items (16 items twice), 42 for English language discrete items (21 items twice) and 28 for English passage-related items, for a total of 112 analyses. Since any of the four terms in each analysis could be statistically significant, there were 448 statistical tests made.

In order to control what might otherwise have been a soaring alpha error level, alpha was set to approximately .05 for the analysis of each of the 5 tryout unit sets (single gender mathematics, mixed gender mathematics, single gender English language discretes, mixed gender English language discretes, and English language passage-related). Available chi-squared tables did not provide sufficient detail to permit selection of the exact critical value necessary to given an $\alpha = .05$ for each of these tables (n items by 4 tests). The best available approximation was the critical value for $\alpha = .0005$. This gives an actual α level of 0.32 for the mathematics tables, of .042 for the English language discrete tables and of .056 for the English passage tables.

In these analyses, degrees of freedom reflect the difference in the number of parameters estimated in the two models. Looking up critical values for four degrees of freedom for the tests of models 5 with 2 and 6 with 5, and two degrees of freedom for the tests of models 8 with 7 and 9 with 8 yielded critical chi-squares of 20.00 and 15.20 respectively.

APPENDIX B

Mathematics Items

Number 60 - Neutral

9. A cyclist rode a bicycle 18 miles before lunch and 3 miles after lunch. It took the cyclist 3 hours of riding to go the total distance. If a second cyclist rides at the same average rate of speed as the first one, how long would it take the second cyclist to travel 28 miles?

A. 2 hours 20 minutes
B. 4 hours
C. 4 hours 28 minutes
D. 8 hours 10 minutes
E. 9 hours 20 minutes

9	1	2*	3	4	5	0
UPP	2	76	13	3	5	1
MID	6	53	31	6	5	0
LOW	2	40	40	9	9	0
TOT	4	59	26	5	6	0

Number 60 - Male

3. Sam rode his bicycle 18 miles before lunch and 3 miles after lunch. It took him 3 hours of riding to go the total distance. If David rides at the same average rate of speed as Sam, how long would it take him to travel 28 miles?

A. 2 hours 20 minutes
B. 4 hours
C. 4 hours 28 minutes
D. 8 hours 10 minutes
E. 9 hours 20 minutes

3	1	2*	3	4	5	0
UPP	1	93	3	2	0	0
MID	2	73	20	3	2	0
LOW	13	39	38	6	6	0
TOT	3	74	18	3	2	0

Number 60 - Female

6. Jean rode her bicycle 18 miles before lunch and 3 miles after lunch. It took her 3 hours of riding to go the total distance. If Paula rides at the same average rate of speed as Sam, how long would it take her to travel 28 miles?

F. 2 hours 20 minutes
G. 4 hours
H. 4 hours 28 minutes
J. 8 hours 10 minutes
K. 9 hours 20 minutes

6	1	2*	3	4	5	0
UPP	1	77	16	0	2	4
MID	4	56	28	7	4	2
LOW	13	50	24	7	4	0
TOT	5	62	22	4	3	3

APPENDIX B (cont.)

Mathematics Items

Number 55 - Neutral

6. A student earned $1,600 last summer. The student spent $224 on a new bicycle, invested $1,000 in a savings certificate, and put the remaining money in a savings account. What percent of the student's earnings did the student put in the savings account?

F. 14.0
G. 20.6
H. 23.5
J. 30.7
K. 37.6

	6	1	2	3*	4	5	O
UPP	9	10	65	3	10	3	
MID	11	11	41	7	31	O	
LOW	21	19	13	23	23	O	
TOT	12	12	44	8	22	1	

Number 55 - Male

6. Gene earned $1,600 last summer. He spent $224 on a new bicycle, invested $1,000 in a savings certificate, and put the remaining money in his savings account. What percent of his earnings did he put in his savings account?

F. 14.0
G. 20.6
H. 23.5
J. 30.7
K. 37.6

	6	1	2	3*	4	5	O
UPP	7	9	69	4	10	O	
MID	17	17	27	14	25	1	
LOW	18	13	18	18	33	O	
TOT	13	13	42	11	20	O	

Number 55 - Female

6. Carmen earned $1,600 last summer. She spent $224 on a new bicycle, invested $1,000 in a savings certificate, and put the remaining money in her savings account. What percent of her earnings did she put in her savings account?

F. 14.0
G. 20.6
H. 23.5
J. 30.7
K. 37.6

	6	1	2	3*	4	5	O
UPP	9	4	70	2	15	O	
MID	15	14	30	13	27	O	
LOW	14	9	25	9	41	2	
TOT	12	10	44	9	25	O	

APPENDIX C

English Usage Discrete Sentences

Number 30 - Neutral

Another group of French film directors incorporated plot lines into their films.

2. F. NO CHANGE
G. about
H. with
J. to

2	1*	2	3	4	5	0
UPP	81	1	10	7	0	0
MID	51	6	26	17	0	0
LOW	29	22	29	20	0	0
TOT	55	7	23	15	0	0

Number 30 - Male

Another Frenchman incorporated plot lines into his films.

16. F. NO CHANGE
G. about
H. with
J. to

16	1*	2	3	4	5	0
UPP	86	0	10	4	0	0
MID	73	9	12	5	1	0
LOW	55	18	22	2	2	0
TOT	73	8	13	4	1	0

Number 30 - Female

Another Frenchwoman incorporated plot lines into her films.

13. A. NO CHANGE
B. about
C. with
D. to

13	1*	2	3	4	5	0
UPP	73	3	15	8	0	0
MID	68	10	13	9	0	0
LOW	50	7	30	13	0	0
TOT	66	7	17	10	0	0

APPENDIX C (cont.)

English Usage Discrete Sentences

Number 31 - Neutral

Most of us thought of cooking as drudgery—
3
something to be hurried through.

3. A. NO CHANGE
B. has thought
C. think
D. having thought

3	1*	2	3	4	5	0
UPP	15	3	81	1	0	0
MID	10	7	81	2	0	0
LOW	5	12	71	12	0	0
TOT	10	7	79	3	0	0

Number 31 - Male

Most of us boys thought of cooking as drudgery—
17
something to be hurried through.

17. A. NO CHANGE
B. has thought
C. think
D. having thought

17	1*	2	3	4	5	0
UPP	21	0	79	0	0	0
MID	29	1	69	1	0	0
LOW	37	2	57	2	0	2
TOT	29	1	69	1	0	0

Number 31 - Female

Most of us girls thought of cooking as drudgery—
8
something to be hurried through.

8. F. NO CHANGE
G. has thought
H. think
J. having thought

8	1*	2	3	4	5	0
UPP	20	0	80	0	0	0
MID	22	3	76	0	0	0
LOW	14	9	68	9	0	0
TOT	20	3	75	2	0	0

APPENDIX C (cont.)

English Usage Discrete Sentences

Number 37 - Neutral

Imagination operates in cooking just as it does in the composition of a poem or a symphony.

9. A. NO CHANGE
B. in a manner similar to that
C. just like they do
D. just as it did

9	1*	2	3	4	5	0
UPP	82	18	0	0	0	0
MID	76	17	2	4	0	0
LOW	41	34	17	7	0	0
TOT	73	20	4	3	0	0

Number 37 - Male

For George, imagination operates in cooking just as it does in the composition of a poem or a symphony.

3. A. NO CHANGE
B. in a manner similiar to that
C. just like they do
D. just as it did

3	1*	2	3	4	5	0
UPP	69	25	3	3	0	0
MID	58	29	9	3	1	0
LOW	37	29	14	10	0	0
TOT	57	28	10	4	0	0

Number 37 - Female

For Janet, imagination operates in cooking just as it does in the composition of a poem or a symphony.

18. F. NO CHANGE
G. in a manner similar to that
H. just like they do
J. just as it did

18	1*	2	3	4	5	0
UPP	82	17	0	1	0	0
MID	76	16	3	5	0	0
LOW	59	38	11	11	2	0
TOT	70	21	4	5	0	0

APPENDIX C (cont.)

English Usage Discrete Sentences

Number 32 - Neutral

The rules in bridge call for formal rounds of bidding, but any other attempt to communicate information concerning the cards a player holds are forbidden.

13. A. NO CHANGE
B. to try and
C. you make to
D. by which they

13	1*	2	3	4	5	0
UPP	84	9	6	1	0	0
MID	57	32	10	1	0	0
LOW	50	34	14	2	0	0
TOT	63	26	10	1	0	0

Number 32 - Male

The rules in bridge call for formal rounds of bidding, but any other attempt to communicate information concerning the cards Ralph and Rich hold are forbidden.

8. F. NO CHANGE
G. to try and
H. you make to
J. by which they

8	1*	2	3	4	5	0
UPP	61	11	22	5	0	0
MID	37	39	17	8	0	0
LOW	27	29	24	20	0	0
TOT	43	28	20	9	0	0

Number 32 - Female

The rules in bridge call for formal rounds of bidding, but any other attempt to communicate information concerning the cards Deb and Sue hold are forbidden.

15. A. NO CHANGE
B. to try and
C. you make to
D. by which they

13	1*	2	3	4	5	0
UPP	71	18	8	2	0	0
MID	45	31	13	10	0	1
LOW	32	30	25	13	0	0
TOT	52	27	13	8	0	0

APPENDIX C (cont.)

English Usage Discrete Sentences

Number 33 - Neutral Single Gender Unit

In bridge, conversation is strictly limited. The rules call for formal rounds of bidding, but any other attempt to communicate information concerning the cards a player holds are forbidden.

5. A. NO CHANGE
B. have to be
C. have been
D. is

	5	1	2	3	4*	5	0
UPP	7	0	1	91	0	0	
MID	15	2	5	78	0	0	
LOW	27	20	15	39	0	0	
TOT	15	4	6	75	0	0	

Number 33 - Male Single Gender Unit

The rules in bridge call for formal rounds of bidding, but any other attempt to communicate information concerning the cards Ralph and Rick hold are forbidden.

19. A. NO CHANGE
B. have to be
C. have been
D. is

	19	1	2	3	4*	5	0
UPP	14	0	3	83	0	0	
MID	32	3	16	49	1	0	
LOW	18	16	22	41	0	2	
TOT	25	5	14	56	0	0	

Number 33 - Female Single Gender Unit

The rules in bridge call for formal rounds of bidding, but any other attempt to communicate information concerning the cards Deb and Sue hold are forbidden.

10. F. NO CHANGE
G. have to be
H. have been
J. is

	10	1	2	3	4*	5	0
UPP	17	0	11	72	0	0	
MID	16	2	13	65	1	0	
LOW	23	9	25	43	0	0	
TOT	19	3	15	62	1	0	

APPENDIX C (cont.)

English Usage Discrete Sentences

Number 33 - Neutral Mixed Gender Unit

The rules in bridge call for formal rounds of bidding, but any other attempt to communicate information concerning the cards a player holds are forbidden.

14. F. NO CHANGE
G. have to be
H. have been
J. is

14	1	2	3	4*	5	0
UPP	11	0	0	89	0	0
MID	27	1	3	70	0	0
LOW	34	6	12	46	2	0
TOT	24	2	4	70	0	0

Number 33 - Male Mixed Gender Unit

The rules in bridge call for formal rounds of bidding, but any other attempt to communicate information concerning the cards Ralph and Rich hold are forbidden.

9. A. NO CHANGE
B. have to be
C. have been
D. is

9	1	2	3	4*	5	0
UPP	17	0	1	81	0	0
MID	24	4	13	59	1	0
LOW	24	7	24	44	0	0
TOT	22	3	11	63	0	0

Number 33 - Female Mixed Gender Unit

The rules in bridge call for formal rounds of bidding, but any other attempt to communicate information concerning the cards Deb and Sue hold are forbidden.

16. F. NO CHANGE
G. have to be
H. have been
J. is

16	1	2	3	4*	5	0
UPP	16	1	7	76	0	0
MID	26	3	18	53	0	1
LOW	35	17	25	22	0	0
TOT	24	5	16	55	0	0

APPENDIX C (cont.)

English Usage Discrete Sentences

Number 36 - Neutral

We were all listening intently when they had began
4
to play.

4. F. NO CHANGE
G. begun
H. began
J. would begin

4	1	2	3*	4	5	0
UPP	3	3	94	0	0	0
MID	5	9	83	2	0	0
LOW	14	12	72	2	0	0
TOT	8	8	84	2	0	0

Number 36 - Male

We were all listening intently when he
had began to play.
H

18. F. NO CHANGE
G. begun
H. began
J. would begin

18	1	2	3*	4	5	0
UPP	0	1	99	0	0	0
MID	2	8	89	2	0	0
LOW	2	11	84	2	0	0
TOT	1	6	91	1	0	0

Number 36 - Female

We were all listening intently when she had began
3
to play.

3. A. NO CHANGE
B. begun
C. began
D. would begin

3	1	2	3*	4	5	0
UPP	0	6	93	1	0	0
MID	4	12	84	1	0	0
LOW	7	13	80	0	0	0
TOT	3	10	86	1	0	0

APPENDIX C (cont.)

English Usage Discrete Sentences

Number 40 - Neutral

Many historians conclude that the ruler was insane; in fact, they consider the ruler the worst of those who fought for self-preservation and territorial expansion.

10. F. NO CHANGE
G. who have been fighting
H. whom fought
J. whom have been fighting

10	1*	2	3	4	5	0
UPP	80	8	10	2	0	0
MID	44	16	36	4	0	0
LOW	27	13	42	17	0	0
TOT	53	13	29	5	0	0

Number 40 - Male

Many historians conclude that the king was insane; in fact, they consider him the worst of the kings who fought for self-preservation and territorial expansion.

5. A. NO CHANGE
B. who have been fighting
C. whom fought
D. whom have been fighting

5	1*	2	3	4	5	0
UPP	87	1	9	3	0	0
MID	68	4	24	4	0	0
LOW	50	10	32	8	0	0
TOT	70	4	21	5	0	0

Number 40 - Female

Many historians conclude that the queen was insane; in fact, they consider her the worst of the queens who fought for self-preservation and territorial expansion.

6. F. NO CHANGE
G. who have been fighting
H. whom fought
J. whom have been fighting

6	1*	2	3	4	5	0
UPP	71	7	16	5	0	0
MID	57	10	27	6	0	0
LOW	33	13	47	7	0	0
TOT	57	10	27	6	0	0

APPENDIX C (cont.)

English Usage Discrete Sentences

Number 48 - Neutral

Convince me it's uncalculated that any evening where I try to make a call, my teenager was beginning a "life-or-death" conversation with a friend.

3. A. NO CHANGE
B. which
C. wherever
D. OMIT the underlined portion.

	3	1	2	3	4*	5	0
UPP	5	6	0	89	0	0	
MID	2	11	5	79	2	1	
LOW	13	24	18	44	0	0	
TOT	5	12	6	76	1	0	

Number 48 - Male

Convince me it's uncalculated that any evening where I try to make a call, my son was beginning what he insists is a life-or-death conversation with his school locker partner.

1. A. NO CHANGE
B. which
C. wherever
D. OMIT the underlined portion.

	1	1	2	3	4*	5	0
UPP	4	17	4	76	0	0	
MID	5	30	4	58	2	0	
LOW	15	27	25	32	0	0	
TOT	6	25	7	60	1	0	

Number 48 - Female

Convince me it's uncalculated that any evening where I try to make a call, my daughter was beginning what she insists is a life-or-death conversation with her school locker partner.

2. F. NO CHANGE
G. which
H. wherever
J. OMIT the underlined portion.

	2	1	2	3	4*	5	0
UPP	0	9	0	90	1	0	
MID	4	16	8	71	1	0	
LOW	8	18	16	58	0	0	
TOT	3	14	7	74	1	0	

APPENDIX C (cont.)

English Usage Discrete Sentences

Number 49 - Neutral

Convince me it's uncalculated that any evening where I try to make a call, my teenager was beginning a "life-or-death" conversation with a friend.

4. F. NO CHANGE
G. had begun
H. has began
J. is beginning

4	1	2	3	4*	5	0
UPP	2	25	15	57	0	0
MID	15	37	17	30	1	0
LOW	11	49	18	22	0	0
TOT	10	35	17	37	0	0

Number 49 - Male

Convince me it's uncalculated that any evening where I try to make a call, my son was beginning what he insists is a life-or-death conversation with his school locker partner.

2. F. NO CHANGE
G. had begun
H. has began
J. is beginning

2	1	2	3	4*	5	0
UPP	1	23	16	60	0	0
MID	5	38	24	33	0	0
LOW	7	38	20	35	0	0
TOT	4	33	21	42	0	0

Number 49 - Female

Convince me it's uncalculated that any evening where I try to make a call, my daughter was beginning what she insists is a life-or-death conversation with her school locker partner.

3. A. NO CHANGE
B. had begun
C. has began
D. is beginning

3	1	2	3	4*	5	0
UPP	0	11	33	56	0	0
MID	6	29	32	33	0	0
LOW	10	40	20	29	0	2
TOT	5	26	30	38	0	0

APPENDIX D

English Usage Passage-Set Items

Number 9 - Neutral

lost-and-found; no one calls to say that they has
stumbled across my pen; my stack of catalogues

9. A. NO CHANGE
B. someone have
C. you have
D. one has
E. he or she has

9	1	2	3	4	5*	0
UPP	0	0	3	8	89	0
MID	2	1	4	12	81	0
LOW	2	4	9	17	68	0
TOT	2	2	4	12	81	0

Number 9 - Male

found; nobody calls to say that they has stumbled
across my billfold; my stack of magazines is only a

9. A. NO CHANGE
B. someone have
C. you have
D. one has
E. he has

9	1	2	3	4	5*	0
UPP	3	0	0	15	82	0
MID	1	1	3	32	63	0
LOW	13	5	8	38	36	0
TOT	3	2	3	28	64	0

Number 9 - Female

found; nobody calls to say that they has stumbled

across my earring; my stack of magazines is only a

9. A. NO CHANGE
B. someone have
C. you have
D. one has
E. she has

9	1	2	3	4	5*	0
UPP	0	1	1	45	52	0
MID	1	4	2	68	25	0
LOW	15	15	10	45	15	0
TOT	3	5	3	58	31	0

APPENDIX D (cont.)

English Usage Passage-Set Items

Number 19 - Neutral

but I
do notice, as I pass the full-length mirror, that my
cheeks are more fulfilled than they were a week ago,
19
and that fitting into my pants seem just a
20
touch tighter.

19. A. NO CHANGE
B. are more fuller
C. are fuller
D. is fuller
E. are more filled

19	1	2	3*	4	5	0
UPP	2	3	89	2	5	0
MID	4	3	81	2	9	0
LOW	9	13	57	2	19	0
TOT	5	5	78	2	10	0

Number 19 - Male

but I do notice, as I pass the full-length mirror,
that my shoulders are more broad than they were a
19
week ago, and that fitting into my trousers seem
20
just a touch tighter.

19. A. NO CHANGE
B. are more broader
C. are broader
D. is broader
E. are more broadened

19	1	2	3*	4	5	0
UPP	4	0	96	0	0	0
MID	15	2	70	4	9	0
LOW	8	5	77	3	8	0
TOT	11	2	79	2	6	0

Number 19 - Female

but I do notice, as I pass the full-length mirror,
that my face is more fulfilled than it was a
19
week ago, and that fitting into my skirts seem
20
just a touch tighter.

19. A. NO CHANGE
B. is more fuller
C. is fuller
D. are fuller
E. is more filled

19	1	2	3*	4	5	0
UPP	4	1	94	0	0	0
MID	7	4	85	1	4	0
LOW	13	27	40	2	17	0
TOT	7	7	80	1	5	0

APPENDIX D (cont.)

English Usage Passage-Set Items

Number 20 - Neutral

but I do notice, as I pass the full-length mirror, that my cheeks are more fulfilled than they were a week ago, and that fitting into my pants seem just a touch tighter.

20. F. NO CHANGE
G. fitting in
H. the fit of
J. that fitting in
K. that

	1	2	3	4	5*	0
UPP	20	3	33	0	44	0
MID	17	16	37	3	28	0
LOW	30	21	30	4	13	0
TOT	20	14	34	3	29	0

Number 20 - Male

but I do notice, as I pass the full-length mirror, that my shoulders are more broad than they were a week ago, and that fitting into my trousers seem just a touch tighter.

20. F. NO CHANGE
G. fitting in
H. the fit of
J. that fitting in
K. that

	1	2	3	4	5*	0
UPP	5	1	27	1	64	0
MID	15	9	36	1	39	0
LOW	26	26	15	5	28	0
TOT	14	9	30	2	45	0

Number 20 - Female

but I do notice, as I pass the full-length mirror, that my face is more fulfilled than it was a week ago, and that fitting into my skirts seem just a touch tighter.

20. F. NO CHANGE
G. fitting in
H. the fit of
J. that fitting in
K. that

	1	2	3	4	5*	0
UPP	8	6	35	0	51	0
MID	26	14	36	3	21	0
LOW	38	27	15	13	7	0
TOT	23	14	33	4	27	0

APPENDIX D (cont.)

English Usage Passage-Set Items

Number 28 - Neutral

28. The age and gender of the person referred to as "I" in this passage are not stated. However, this person is probably:

F. a male child.
G. a male adult.
H. a female child.
J. a female adult.
K. an adult (gender not indicated).

28	1	2	3	4	5*	0
UPP	0	15	2	30	54	0
MID	1	17	1	35	44	1
LOW	2	13	0	32	53	0
TOT	1	16	1	33	48	1

Number 28 - Male

28. The age and gender of the person referred to as "I" in this passage are not stated. However, this person is probably:

F. a male child.
G. a male adult.
H. a female child.
J. a female adult.
K. an adult (gender not indicated).

28	1	2*	3	4	5	0
UPP	3	58	0	12	27	0
MID	1	44	0	20	33	1
LOW	0	31	0	31	33	5
TOT	2	46	0	19	32	1

Number 28 - Female

28. The age and gender of the person referred to as "I" in this passage are not stated. However, this person is probably:

F. a male child.
G. a male adult.
H. a female child.
J. a female adult.
K. an adult (gender not indicated).

28	1	2	3	4*	5	0
UPP	1	0	0	94	3	1
MID	0	0	0	83	11	1
LOW	2	2	2	72	17	2
TOT	1	0	0	88	10	1